Iterative Data Modelling: from Teaching Practice to Research Method

Pim van Bree
pim@lab1100.com
LAB1100, The Netherlands

Geert Kessels
geert@lab1100.com
LAB1100, The Netherlands

Introduction

Data modelling is an essential process of almost any digital humanities project (Flanders and Jannidis, 2015). Whether texts, images, or any other form of data is mapped or analysed, a model has to be conceptualised that describes the data and forms the bedrock of the application that contains or analyses the data.

Since data modelling in the humanities is largely perceived as an epistemological process, rather than an ontological process, there is a tension between the way in which material and knowledge presents itself and the manner in which material and knowledge can be described on a generalised or abstracted level. As Flanders and Jannidis (2015: 236) have pointed out: "Some of the most fertile and urgent areas of digital humanities research involve the question of how to develop data modeling approaches that accommodate both the self-reflexivity required by humanities research and the actionability and computational clarity required by the digital domain."

In this paper we reflect on a data modelling approach that has proven to be an effective teaching practice as well as a useful research method. The iterative data modelling approach we put forward focuses on a continuous shift between three levels of data modelling: conceptual level, logical level, and interface level. We have found that this approach provides students and scholars in the humanities ("scholars") with the skills they need to translate their body of data or research question into an operational process that produces rich (inclusive; fuzzy and uncertain) and complex (advocate divergent classes) actionable datasets. It is important to note that even though it is useful when a scholar can develop their own data model for computer-aided analytical purposes, we should not underestimate the learning curve this new skillset requires.

This paper focuses on experiences we gained from data modelling practices in the humanities aimed at developing a relational database. We draw on the results of over 20 courses and workshops for scholars we have held in the past three years on developing data models and using database applications. These insights are also informed by the continuous development of the research environment nodegoat, developed by the authors of this paper, and the scholarly collaborations resulting therefrom.

Challenges

Most scholars do not perceive their material or knowledge as 'data' (Posner, 2015). Once a scholar has accepted that lists of people, statements, and ideas can also be seen as data and that we do not necessarily need to be able to count with them, it becomes clear that their material or knowledge can be modelled as well. Like the analog card catalog, a database helps to store data properly and sustainably. This allows us to filter and query the data. We can then also create networks and analyse relationships. It is important to note that vagueness, uncertainty, and incompleteness can be incorporated in a data model.

To allow scholars to operationalise the data modelling process, three levels of a data model have to be studied. The conceptual level, the logical level, and the interface level describe the data at hand, each in its own way. Here, it is necessary to reflect critically on hidden assumptions in the choice of entity types and classifications (Erickson, 2013). An iterative data modelling approach is largely research driven, although existing standards could be used as well. By asking scholars to operationalise their own data model, rather than using or implementing a pre-existing model, they get acquainted with the complexities and granularities that operationalising a data model entails.

The interface challenge - how to operate a database application? - is very important. We see the translation of the data model into an actual database as a vital step to get a good understanding of the data modelling process. We prefer to do this with a database application that has a graphic user interface to be able to iterate quickly and to easily compare data models.

Teaching Practice

The participants in our data modelling workshops ranged from undergraduates to established scholars. These workshops were either in the format of intensive one day workshops or stretched over multiple events in the course of months. During a workshop, we first addressed the aforementioned challenges to show that the challenges participants face are not new and that we can critically reflect on them. Secondly, we did collective exercises to give participants an understanding on how data models and database applications work.

When we then asked them to conceptualise a data model based on their own research question, most participants did not know where to start. The reason for this seemed to be twofold. First, they were unable to process new information regarding data modelling and the database application into an operational process. Since most of the participants were trained to conduct research with a syntagmatic dimension in mind, a linear text, it was hard to execute a research process that leads to a paradigmatic dimension, a database (Manovich, 1999). Secondly, since they were invested in the complex and unique aspects of their research project, they were unable to operationalise a coherent model while keeping relevant variables and complexities in mind (Beretta, 2016).

To overcome this, we introduced an iterative approach that took them back and forth from their research question to a partial conceptual data model, to a partial logical model, and to a partial functional database application. This process helped them to first understand how to translate one class of information to a single, non-relational, data table. Once they could process basic typed values (strings/numbers), they started to work with texts, images, dates, and locations. These steps informed participants on the transformation of a conceptual idea into a table with fields in a database application. They first focused on the basics, the finite, while leaving growing complexity, the infinite, to next iterations: creating additional data tables and constructing relationships between them. After these practical questions had been tackled, attention was shifted towards uncertain data, fuzzy data, and the question on using existing standards for a data model.

In literature on data modelling processes, a distinction is made between the conceptual/logical level and the level of the application. A data model should be portable and not dependant on one application (Flanders and Jannidis, 2015). However, this does not mean that the conceptual/logical level may not be informed by the application while teaching data modelling practices. The feedback loop between these different levels has proven to be an essential step in helping scholars understand how their own research project can be translated into a data model and a functional database application.

Research Method

The iterative data modelling approach is also of value as a research method. The aforementioned distinction between the conceptual/logical level and the interface level works well when the data for a data model is complete and unambiguous and the process in which the data model plays a role is completely mapped out. Obviously, these variables rarely hold true for research projects in the humanities.

Oftentimes the data model does not correspond with data at hand. First, a data model may ask for data that is not there for the majority of data objects. Secondly, data may be too vague to fit typed fields defined in a data model. Thirdly, a data model may lead to a research process that is too time consuming due to its level of detail. In all these cases, revisions of the data model are needed in order to continue the research process.

Instead of smoothing out irregularities in the data by simplifying the data model, the model should be adjusted to reflect the existing complexities, vagueness, and uncertainties. As Rawson and Muñoz (2016) have stated, scholars should "see the messiness of data not as a block to scalability but as a vital feature of the world which our data represents and from which it emerges." We encourage scholars to include these data driven practices into their data model and have developed various strategies and features, such as 'reversed classification', to allow them to do this in nodegoat (van Bree and Kessels, 2014; van Bree and Kessels, 2017).

With the iterative methodology applied in nodegoat, we have facilitated research projects in the range of: disambiguation of Babylonian letters, questions of provenance and intertextuality in medieval manuscripts, creation of a multi-sourced 19th century context of conference attendance on social issues, mapping structures of violence in 1965 Indonesia, and documentation of an actor-network towards an encyclopedia of romantic nationalism.

An iterative data modelling approach allows scholars to enrich their data model during the research process. While a scholar may first want to use their own model, this can later be transformed or mapped to existing standards, like CIDOC-CRM, or semantic web standards. The data itself may be enriched by adding external identifiers such as VIAF identifiers or identifiers to other linked open data resources. This last point is important when data is published as an actionable dataset online (Berners-Lee, 2006).

Conclusion

In this paper we have set out to describe an iterative data modelling approach that helps scholars become confident in modelling their data and that functions as a research method for database development in the humanities. We have argued how a continuous shift between three levels of data modelling helps to conceive actionable datasets and establishes a framework for dealing with the complexities associated with humanities research.

Bibliography

Beretta, F. (2016). From Index Cards to a Digital Information System: Teaching Data Modeling to Master's Students in History. In *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, pp. 132-135.

Berners-Lee, T. (2006) Linked Data. Retrieved from https://www.w3.org/DesignIssues/LinkedData.html.

Bree, P. van and Kessels, G. (2014) Reversed Classification [Blog Post]. Retrieved from https://nodegoat.net/blog.s/5/reversed-classification.

Bree, P. van and Kessels, G. (2015) “Mapping memory landscapes in nodegoat” in: *Social Informatics,* ed. L.M. Aiello and D. McFarland (Lecture Notes in Computer Science 8852), pp 274--278, New York : Springer International.

Bree, P. van and Kessels, G. (2017) Formulating Ambiguity in a Database [Blog Post]. Retrieved from https://nodegoat.net/blog.s/21/formulating-ambiguity-in-a-database.

Erickson, A. T. (2013). Historical Research and the Problem of Categories. In: Dougherty, J. and Nawrotzki, K. (eds), *Writing History in the Digital Age*. Ann Arbor: University of Michigan Press, pp. 133-145.

Flanders, J. and Jannidis, F. (2015) Data Modeling, in *A New Companion to Digital Humanities* (eds S. Schreibman, R. Siemens and J. Unsworth), John Wiley & Sons, Ltd, Chichester, UK. doi: 10.1002/9781118680605.ch16

Manovich, L. (1999). Database as a symbolic form. *Millennium Film Journal*, 34 (Fall)

Posner, M., (2015) Humanities Data: A Necessary Contradiction [Blog Post]. Retrieved from http://miriamposner.com/blog/humanities-data-a-necessary-contradiction/.

Rawson, K. and Muñoz, T. (2016) Against Cleaning. Retrieved from: http://www.curatingmenus.org/articles/against-cleaning.